2К

УДК 004.89:004.93

З.Ю. Кулиева
Институт информационных технологий НАН Азербайджана, г. Баку, Азербайджан
depart15@iit.ab.az

Экспертная система поддержки перевода

В данной статье приводятся результаты экспериментальной реализации морфологического процессора в составе экспертной системы поддержки перевода, представляющей собой экспериментальную двуязычную систему МП с английского языка на азербайджанский. Дается программное описание автоматического словаря как основного элемента ЭСПП и взаимодействие автоматического словаря с базой знаний системы. Представлена методика составления продукционных правил базы знаний, основанная на сочетаемости слов друг с другом в процессе образования их грамматических форм.

Экспертные системы в обработке лингвистических знаний

Современные представления о морфологии в теоретическом языкознании, опирающиеся на теорию интерпретации, на типологию, на огромный фактический материал конкретных языков, создают фундамент для построения систем морфологических знаний. Такие системы не только накапливают информацию о языке в некотором внутреннем формате, но и используют ее для распознавания и синтеза текстов (в режиме верификации). Усовершенствование редакторов текста для конкретного языка тогда можно рассматривать как результат такой оптимизации этих систем, когда из всего богатства универсальных средств отбираются только процедуры и параметры, существенные для конкретного языка. Как известно для сбора, хранения и обработки информации больших объектов данных наиболее распространенной технологией являются базы данных (БД), однако последние не позволяют структурировать хранящиеся в них данные на основе тех отношений, которые существуют между фактами непосредственно в реальной среде. Причем эти отношения должны отражать существенные связи объекта, т.е. позволяет лингвисту на основе своих обновленных сведений давать анализ произвольному выражению объектного языка и синтезировать – при необходимости – словоформы конкретной лексемы, коль скоро задан требуемый набор грамматических категорий.

ЭС, являясь крупным достижением современной вычислительной техники и методов искусственного интеллекта, представляет собой специализированную компьютерную систему, способную к накоплению и обобщению опыта высококвалифицированных экспертов, и моделирует рассуждения последних в некоторой определенной области, используя для этого базу знаний (БЗ), содержащую факты и правила из этой области и некоторую процедуру логического вывода. В последние годы разработка экспертных систем, содержащих различные знания о языке, нашла широкое распространение. Так, например, можно отметить экспертную систему машинного перевода, основанную на базе морфологических и синтаксических знаний, разработанную на основе двухуровневой модели [1], позволяющую без дополнительного программирования создавать описания морфологической структуры естественного языка и списки процедур. Последняя дает возможность использовать морфологические процедуры в при-

кладных программах на языках Си, Пролог и др. Далее представляет интерес ЭС морфологических знаний, позволяющая провести морфологическую интерпретацию текста (истолкование словоформ), на основе постоянно обновляющихся знаний.

Экспериментальная реализация морфологического процессора выполнена в Институте Информационных технологий при НАНА в виде двухуровневой модели английского и азербайджанского языков. Реализация для некоторых других языков потребует введения в исходную модель дополнительных возможностей. Для данной пары языков на данном этапе в рамках экспертной системы поддержки перевода (ЭСПП) довершены описания отдельных подсистем морфологического строя, такие, как системы главных частей речи в сочетании с предлогами, частицами, вспомогательными элементами.

Структура ЭСПП

ЭСПП реализована на базе программы Delphi 7, применяемой для создания систем управления базами данных и знаний. Данная экспертная система работает в двух режимах: перевода фразы (словосочетаний на основе морфосинтаксического анализа и пословного перевода, базирующегося только на морфологическом анализе). Второй режим используется как аварийное средство, когда нельзя построить синтаксическую структуру и выполнить перевод в полном режиме. Объем словаря в реализованной версии составляет 2000 входов на каждый язык в словаре комбинированного типа.

Для реализации морфосинтаксического анализа написан комплекс программ, включающий морфосинтаксический анализатор, позволяющий описывать грамматическую информацию. Создан перечень морфосинтаксических и семантических признаков для пересекающихся классификаций языка; информация о признаках включена в словарные статьи комбинированного словаря английского/азербайджанского языков.

ЭСПП является системой поддержки обучения и перевода на основе базы знаний. База знаний создана для конкретной предметной области, с включением в словарь слов нейтральной лексики. Система включает в свой состав различные объекты и связывающие их правила. Первоначальное создание словаря на основе нейтральной лексики необходимо для проверки работы морфонематических, морфологических и синтаксических правил. При определенных параметрах, таких, как вполне определенная техническая тематика, поверхностный словарь и поверхностно структурная грамматика, использование анализатора может повысить эффективное обучение перевода, одновременно представляющие возможность обучения выбранным рабочим языкам.

ЭСПП работает в настоящее время в экспериментальном режиме, и в связи с этим ее можно охарактеризовать как экспериментальную двуязычную систему МП, которая использует полную морфологию и поверхностный синтаксис рабочих языков.

Процесс перевода с одного языка на другой сводится к преодолению расхождений между языками. В ЭСПП межязыковые расхождения частично снимаются на каждом из этапов анализа, и в основном на этапе трансфера. Трудность преодоления расхождений обусловлена тем, что между значимыми элементами английского и азербайджанского языков трудно найти взаимное соответствие. Одни и те же близкие значения могут кодироваться различными средствами – морфологическими, лексическими и синтаксическими.

Например, расхождение в переводе идиоматических выражений, которыми заполнен лексикон английского языка «when in Austria» – *будучи в Австрии*, или же «the weather permitting» – *если погода позволит.*

ЭСПП представляет собой языковую модель переводных соответствий, работающую на основе трансформационной грамматики с элементами грамматики непосредственно составляющих и конечных автоматов. Хотя, как было указано выше, лингвистическое обеспечение ЭСПП логически отделено от алгоритмов, рассматривать его удобно в алгоритмическом порядке, то есть в порядке включения его отдельных компонентов в процессе перевода.

АС как база данных и один из основных компонентов ЭСПП

Автоматический словарь в составе экспертной системы представляет собой хранилище информации, используемое для обработки текста на основе знаний, представленных в виде трансформационных правил распознавания и порождения грамматических, фонетических и семантических явлений языка.

Автоматический бинарный словарь разрабатывается как часть интегрированной системы перевода и используется для выполнения следующих задач:

– служит основным инструментом поиска (установления) лексических переводных эквивалентов в ЭСПП;

– для работы в диалоговом режиме словарь интегрирован в общую лексикографическую базу ЭСПП и является основной информативно-справочной базой;

– в ЭСПП, как в одной из систем автоматической обработки текста, АС служит источником грамматической информации, необходимой для работы алгоритмов автоматического морфологического и синтаксического анализов, а также для работы алгоритмов лемматизации и правил базы знаний. Последние обеспечивают работу словаря при выполнении любой из названных функций в любой парадигматической формой слова.

В табл. 1 приводится краткое программное описание АС в ЭСПП по следующим критериям:

Таблица 1 – Программное описание АС в ЭСПП

Критерий	Описание
Тип словаря	Автоматический двуязычный словарь в экспертной системе машинного перевода
Рабочая среда	Delphi 7 Studio, Borland
Установка программы	Программа установки словаря предоставляет режим установки на локальный диск
Языки, входящие в состав системы	В состав системы входит англо-азербайджанский словарь
Возможности поиска переводов	Предусматриваются следующие возможности поиска переводов: морфологический (для любой формы слова и запросы, в соответствии с имеющимися лексико-грамматическими данными)
Возможность работы с некоторыми тематиками	В состав системы входит единый словарь, из которого на экран выдаются оптимальные переводы
Ввод новых словарных статей	Новые словарные статьи вводятся либо вручную, либо из текстового файла в режиме импорта
Метод поиска в словаре	Морфологический и синтаксический анализ текста оригинала, использование переводного словаря и словаря исключений

Наличие словарной статьи	Информация в словаре структурирована в виде словарных статей, каждая из которых содержит оригинал, перевод, грамматическую информацию
Разграничение однонаправленных словарных статей	Словарная статья может быть доступна при переводе с английского на азербайджанский язык
Использование лингвистической схемы при обработке запроса	Производится для всех статей в словаре как для переводного, так и для словаря исключений
Работа с фразами (морфологическая обработка и перевод)	Производится морфосинтаксическая обработка фраз
Выделение различных переводов на экране	Переводы разделяются в меню по частям речи, значениям и по фразам
Результаты работы программы	На выходе может быть получен построчный перевод текста запроса
Наличие дополнительных полей	В любую статью можно ввести комментарии и примеры
Удобное и быстрое пополнение словаря	Добавление новой словарной статьи может происходить автоматически и обеспечивает последующую доступность статьи во всех режимах поиска
Количество пополняемых морфологических классов для английского языка	Система позволяет осуществлять ввод словоформ по 65 различным типам склонений и спряжений английского языка, включая некоторые виды исключений
Количество пополняемых морфологических классов для азербайджанского языка	Система содержит 70 классов азербайджанского языка, включая классы с чередованием, выпадением гласных и т.д.
Возможность добавления словаря исключений	Существует возможность включения списков исключений (например, неправильных глаголов)
Автоматическое указание названия части речи в меню переводов	Название части речи выводится после морфологического анализа
Возможность пропуска непереводимого слова из текста запроса	При формировании построчного перевода неизвестные системе слова пропускаются
Максимальное число обрабатываемых слов во фразе	Фразы из 2-значимых слов
Общее количество словарных входов	Общее количество статей по всем тематикам более 2000
Количество статей общей лексики	1500 словарных статей
Объем словаря на жестком диске	170 Кб
Предпосылки для дальнейшего усовершенствования словаря	Формирование построчного перевода текста запроса, возможность просмотра всех фраз в словаре, содержащих данное слово, увеличения объема словаря, применение дополнительных тематик

База знаний ЭСПП

База знаний содержит информацию, необходимую для решения задач требуемого типа, в виде правил и фактов. Механизм вывода представляет собой общий алгоритм решения задач, реализуемый, как правило, в виде интерпретатора. Применение его к базе знаний о конкретной предметной области, задаваемой экспертом, и к данным о текущей ситуации, задаваемой пользователем, дает решение требуемой задачи. Интерфейс с пользователем предназначен для взаимодействия с ним во время решения задачи, и в зависимости от типа задачи, может использовать средства анализа фраз на естественном языке, выбора меню, графического ввода и вывода.

При создании ЭСПП необходимо учитывать следующие аспекты хранения и переработки знаний:

– содержание знаний;
– репрезентация знаний, форма хранения, предназначенная для эффективного поиска, переупорядочения и модифицирования;
– формализация объектного знания (в нашем случае – формализация знаний о языке-объекте);
– переработка текстов как последовательности слов;
– интерпретация синтаксических структур и перевод их в лексическое представление при учете контекста;
– представление знаний внутри самой системы;
– разработка вспомогательных средств для формализации, хранения и поиска знаний при обработке показаний экспертов.

База знаний осуществляет работу механизма ЭСПП на основе трансформационной (порождающей) грамматики, представляющей собой систему правил, экспериментальным образом приписывающую предложениям структурные описания.

Любая трансформационная грамматика имеет в своем составе морфосинтаксический, морфонологический и семантический компонент. Морфосинтаксический компонент определяет бесконечное множество абстрактных формальных объектов, каждый из которых включает в себя всю информацию, существенную для одной интерпретации конкретного предложения.

Морфонологический компонент определяет фонетическую форму предложения, порождаемого синтаксическими правилами. Он соотносит структуру, порождаемую синтаксическим компонентом, с фонетически репрезентированным сигналом. Семантический компонент определяет семантическую интерпретацию предложения. Он соотносит структуру, порождаемую морфосинтаксическим компонентом, с определенной семантической репрезентацией.

Следовательно, морфосинтаксический компонент грамматики должен указывать для каждого предложения глубинную структуру, обусловливающую его семантическую интерпретацию и поверхностную структуру, которая определяет его фонетическую интерпретацию.

Основополагающей идеей трансформационной грамматики является идея о том, что поверхностная структура задается неоднократным применением определенных формальных операций, называемых «грамматическими трансформациями», к объектам элементарного вида. База морфосинтаксического компонента – система правил, порождающая конечное множество базовых цепочек, каждое из которых имеет связанное о ней структурное описание, называется базовым показателем структуры

составляющего. Эти базовые показатели являются элементарными единицами, составляющими глубинные структуры. В основе предложения лежит последовательность базовых показателей, каждый из которых порождается базой синтаксического компонента. Общий смысл предложения зависит не только от смысла его слов, но и от синтаксической структуры предложения. Синтаксическая структура предложения – это совокупность сведений о связях между его словами и словосочетаниями.

Виды правил базы знаний ЭСПП

В качестве единиц хранения – «элементарных знаний» – выступают не только декларативные сведения, но и меняющиеся от эксперта к эксперту предписания – продукционные правила, что делать при том или ином состоянии обследуемого объекта. Задача лингвиста – формулировка лингвистических правил – правил языка и речи. Моделирование лингвистических правил как базового набора предопределяет соответствующую архитектуру экспертной системы.

В исследовании составление правил базы знаний основывается на методике сочетаемости слов друг с другом в процессе образования их грамматических форм, иными словами распознавание тех или иных грамматических форм слов в предложении и выявление их морфологической и синтаксико-семантической принадлежности возможно осуществить в рамках сочетаний этих словоформ друг с другом. Синтаксическое разделение предложения на словосочетания с внутренними подчинительными связями, редко согласованием и примыканием предопределяет работу синтаксического блока анализа.

Знания ЭСПП представлены набором продукционных правил, каждое из которых состоит из: антецедента (условия) и консеквента (результата) [2]. На простом языке пользователя правило состоит из правой и левой части. Знания ЭСПП представляют собой комплекс правил унификационной грамматики, которая включает в свой состав элементы грамматик разных видов, таких, как: контекстно-свободная грамматика (КСГ), обеспечивающая морфологический анализ и синтез и являющаяся основой анализаторов, цепочечная грамматика (ЦГ) и грамматика непосредственно составляющих (ГНС), обеспечивающие синтаксический анализ и синтез.

Так, элементы КСГ формализуют описание языковой модели как формальной грамматики с конечным числом состояний. Элементы ЦГ фиксируют порядок следования объектов цепочки формально-языковой модели, то есть линейные структуры предложения формальной языковой модели, заданные в терминах грамматических классов слов. В ЭСПП применяется стратегия анализа «слева направо»: перебор слов, проверка условий, наличие или отсутствие изменений по условиям и добавление недостающих элементов формально представляют собой компьютерную реализацию грамматики с конечным числом состояний или КСГ, построенной на ЦГ. В базе знаний системы ЭСПП синтаксическая структура предложения может быть представлена:

1) деревом синтаксического согласования или подчинения линейных узлов, т.е. слова в предложении находятся в несимметричных отношениях друг к другу (одни слова подчиняют себе другие), а формальное подчинение состоит в том, что одно слово определяет грамматическую форму другого;

2) деревом синтаксического подчинения или просто деревом подчинения, заданным на множестве словоформ предложения.

С учетом морфосинтаксических и семантических признаков английского и азербайджанского языков правила в базе знаний представлены следующими видами сочетаемости:

– существительное в им.п. + существительное в притяжательном падеже (категория принадлежности);
– предлог + артикль + существительное;
– предлог + существительное (падежные эквиваленты в азербайджанском языке);
– существительное + to be (категория сказуемости существительных);
– существительное + ед.ч/мн.ч + глагол (временные формы глагола страдательного и действительного залога).

Собственно-синтаксические правила разделяются на именную, предложную и глагольную группы. В формальном описании правила можно разделить на:

– правила распознавания;
– правила порождения;
– правила подстановки.

Разработанная экспертная система поддержки перевода (ЭСПП) обладает следующими особенностями:

1) ориентирована на конкретную область экспертизы, в данном случае на перевод текстов с английского языка на азербайджанский язык;

2) способна делать выводы из посылок: четко сформулированные условия правил предопределяют качества, присущие конечным автоматам;

3) способна пополняться по ходу и в результате работы, охватывая все более широкие наборы знаний;

4) основана на наборе правил, в том числе – на практических правилах, формулируемых экспертом-человеком;

5) обладает практической ценностью.

Литература

1. Гельбух А.Ф. Эффективно реализуемая модель морфологии флективного естественного языка: Автореф. дис… к.ф.н. / АНРФ., ВИНИТИ. – М.,1994.
2. Демьянков В.З. Морфологическая интерпретация текста и ее моделирование. – Из-во Московского университета, 1994. – Режим доступа: http://www.infolex.ru.

Z.Y. Kuliyeva

Translation Support Expert System

In the article there are shown the results of experimental realization of morphological processor within translation support expert system which represents experimental bilingual MT system from English into Azerbaijani. Program description of automatic dictionary as a basic element of translation support expert system and interaction of automatic dictionary with knowledge base of the system are shown herein. Compiling methods of generative rules based on the combinability principle of words with each other in the process of their grammatical forms generation are represented.

Статья поступила в редакцию 18.07.2008.